独立成分分析

赤穂昭太郎

1　独立成分分析概要

カクテルパーティ効果　n 人の話者が話をしているのを m 個のマイクで取り，それぞれの話者の話を聞き分ける．これを独立成分分析 (ICA = independent component analysis) とか，BSS (= blind source separation) とかいう．

記号の定義　話者の発話信号: $\boldsymbol{s}(t) = (s_1(t), \ldots, s_n(t))^\top \in \mathbb{R}^n$
マイクで取った信号: $\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))^\top \in \mathbb{R}^m$
時間遅れがないと仮定し[1]，i さんの発話 $s_i(t)$ は $a_{ji} \in \mathbb{R}$ に比例してマイク j に届くとすると[2]，

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) \tag{1}$$

と書ける．$(A = (a_{ji}))$

さらなる仮定　話者とマイクの数は同じ，すなわち，$n = m$ とする[3]．また，A は full rank とする．

問題　$t = 1, 2, 3, \ldots, T$ までの観測，$\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)$ をもとに，$\boldsymbol{s}(1), \ldots, \boldsymbol{s}(T)$ を推定する (ただし $T > n$)．

[1]時間遅れのある場合はたたみこみで書ける．詳しくは 7 節
[2]非線形の場合は非線形 ICA という話があるが十分研究されていないので省略
[3]マイクの方が多い場合は因子分析や stiefel manifold という話，少ない場合は sparse coding を仮定するというように問題の性質がかなり違ってくる．

規準　A を知っていれば $\boldsymbol{s}(t) = A^{-1}\boldsymbol{x}(t)$ となるが，それは知らないとしよう．その代わり，$\boldsymbol{s}(t)$ を確率過程の見本過程と見て，$s_1(t), \ldots, s_n(t)$ が互いに独立であると仮定する[4]． すると，それを満たすように A^{-1} の推定値 W を見つけることができる可能性がある[5]．それを用いて

$$\boldsymbol{y}(t) = W\boldsymbol{x}(t) \tag{2}$$

により $\boldsymbol{s}(t)$ の推定値 $\boldsymbol{y}(t)$ が得られる．

簡単のため $\boldsymbol{s}(t)$ は i.i.d. (independently identically distributed) であり，密度関数 $p_{\boldsymbol{s}}(\boldsymbol{s})$ をもつとしよう[6]．

独立性：i.i.d. なので，t を外して書くと，$s_1, \ldots, s_n$ が独立というのは，

$$p_{\boldsymbol{s}}(\boldsymbol{s}) = \prod_{i=1}^{n} p_{s_i}(s_i) \tag{3}$$

ただし， $p_i(s_i)$ は s_i の周辺密度 $p_{s_i}(s_i) = \int p_{\boldsymbol{s}}(\boldsymbol{s}) d\boldsymbol{s}_{-i}$ とする[7].

不定性　独立性の規準は基本的に成分の順番と成分毎の線形変換に関して不変である．

任意の置換 $i_1, \ldots, i_n$ に関して，0 でない $a_1, \ldots, a_n$ を取ると，$a_1 s_{i_1}, \ldots, a_n s_{i_n}$ もやはり独立である．

また，正規分布の線形結合はやはり正規分布になるので，正規分布は高々 1 個だけと仮定する．

逆に，正規分布が 1 個だけであれば独立成分分析が well-defined になるというのが以下の定理である．

定理 1　x_1, x_2 を互いに独立な実確率変数 で，

$$y_1 = a_{11}x_1 + a_{12}x_2, \quad y_2 = a_{21}x_1 + a_{22}x_2, \tag{4}$$

とおく． y_1, y_2 が独立なら，この変換は自明な変換であるか，あるいは，これらはすべて正規分布である．

---

[4]通常確率変数 S を大文字で書いて，その見本過程や実現値 (サンプル) は小文字で書くのが確率論の通例だが，見づらいのでここではすべて小文字で書き，紛らわしい場合は明示する．

[5]それが見つけられるための条件は追って説明していく

[6]これは強すぎる仮定であり，実際には，t に関する独立性は強過ぎで，とりあえず強定常であればよい． 実際にはエルゴード性が満たされればよい．

[7]$\boldsymbol{s}_i$ は s_i を除いた $\boldsymbol{s}$ の成分 $(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ の省略記法

この定理から以下の定理が示される．

定理 2 (同定可能定理) 独立な信号 $\boldsymbol{s} = (s_1, \ldots, s_n)^\top$ のうち正規分布は高々一つであるとする．観測信号 $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ を線形変換 $\boldsymbol{y} = W\boldsymbol{x}$ によって変換した信号 $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ がどの二つをとっても独立なら，全体は独立である． また，$\boldsymbol{s}$ と $\boldsymbol{y}$ は自明な変換 *(*置換と各成分の定数倍*)* を除いて一致する．

略証　簡単のため，密度関数が 2 階微分可能として証明する．

x_1, x_2, y_1, y_2 の密度関数を便宜上指数の肩の上にのせた $\exp(\phi_1(x_1))$, $\exp(\phi_2(x_2))$, $\exp(\psi_1(y_1))$, $\exp(\psi_2(y_2))$ の形で書くと，独立性の条件から，x_1, x_2 の同時密度と y_1, y_2 への同時密度は

$$p(x_1, x_2) = \exp(\phi_1(x_1) + \phi_2(x_2)), \qquad q(y_1, y_2) = \exp(\psi_1(y_1) + \psi_2(y_2)), \tag{5}$$

となる．

任意の可測集合 B について， $\boldsymbol{y} = A\boldsymbol{x}$ のとき，

$$\begin{aligned} \Pr[\boldsymbol{x} \in B] &= \int_B p_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} = \int_{B'} p_{\boldsymbol{x}}(A^{-1}\boldsymbol{y}) \frac{d\boldsymbol{y}}{|\det A|} \\ &= \Pr[\boldsymbol{y} \in B'] = \int_{B'} p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y} \end{aligned} \tag{6}$$

なので (ただし，B' は B を A で線形変換した集合)，

$$p_{\boldsymbol{y}}(\boldsymbol{y}) = \frac{p_{\boldsymbol{x}}(A^{-1}\boldsymbol{y})}{|\det A|} \tag{7}$$

である．

x_1, x_2 から y_1, y_2 への変換は正則であるとしてよいので，その Jacobian を

$$c = \left| \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \right| \tag{8}$$

と置くと，$p(x_1, x_2) = cq(y_1, y_2)$, つまり，

$$\phi_1(x_1) + \phi_2(x_2) = \psi_1(a_{11}x_1 + a_{12}x_2) + \psi_2(a_{21}x_1 + a_{22}x_2) + \log c. \tag{9}$$

これを x_1, x_2 で順に微分すると，

$$a_{11}a_{12}\ddot{\psi}_1(a_{11}x_1 + a_{12}x_2) + a_{21}a_{22}\ddot{\psi}_2(a_{21}x_1 + a_{22}x_2) = \\ a_{11}a_{12}\ddot{\psi}_1(y_1) + a_{21}a_{22}\ddot{\psi}_2(y_2) = 0. \tag{10}$$

$a_{11}a_{12} = 0$ のときは，正則性から自明な変換しかあり得ない．一方，$a_{11}a_{12} \neq 0$ のときは，$C = \{(x_1, x_2) \mid y_1 = \text{const.}\}$ という集合を考えると，正則性から C 上の (x_1, x_2) で定義される y_2 はすべての実数を取りうる．従って，C 上で考えれば，すべての y_2 に対して

$$\ddot{\psi}_2(y_2) = \text{const.} \tag{11}$$

同様に，$\ddot{\psi}_1(y_1) = \text{const.}$ が言える．微分方程式を解くと，

$$\psi_i(y_i) = \alpha_i y_i^2 + \beta_i y_i + \gamma_i \tag{12}$$

となる． 実確率変数について $\alpha_i < 0$ でなければならないので，これは正規分布にほかならない．

その他の応用　「観測信号から，その中に含まれている複数の独立成分を取り出す」という問題は一般的であり応用も広い．

例:

1. ノイズ除去：音響信号・画像信号からノイズを除く
2. 生体信号解析：脳信号 (fMRI, 脳波など) から，脳活動の独立な成分を探す．
3. 脳の理解：脳の一次視覚野では，入力画像に含まれる独立成分を抽出しているらしい．

2　主成分分析と回転の自由度

以下簡単のため，信号の平均は 0 と仮定する．必要ならば全体から平均（の推定値）を引いてやればよい．

独立性の必要条件として，無相関性： $E[s_i s_j] = 0 (i \neq j)$ がある．

多変量解析の一手法として知られる主成分分析と関係がある．

$x = [\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)]^\top$ とし，その特異値分解を

$$x = UDV^\top \tag{13}$$

とする．x は $T \times n$ の full-rank 行列とすると，U は $T \times n$ で，$U^T U = I$, V は n 次直交行列．D は n 次対角行列で，その対角成分は特異値と呼ばれ，正の値を取る． 便宜上降順に並んでいるとする．

x に右から V をかけてやれば，UD となり，この各列は直交，つまり無相関にできる[8]．

主成分分析では低次元化が主眼だが，ここでは単に無相関化の手段としてとらえる． さて，UD にさらに右から D^{-1} とすれば，成分の正規化ができるが，実は U は一意的ではない．U の右側から任意の直交行列 W をかけてやると，やはりこれも x の特異値分解を与える．

従って，無相関性だけでは独立成分を抽出するのに十分ではない．

3 独立性の規準

基本的な方針は，何か独立性を測る目的関数を考えてそれを最適にするパラメータを求めるということである．

密度関数に基づく規準 独立性の素朴な定義ははじめにも述べたように，

$$p_{\boldsymbol{y}}(\boldsymbol{y}) = \prod_{i=1}^{n} p_{y_i}(y_i) \tag{14}$$

なので，独立信号は右辺と左辺の距離が 0 になるところを探す問題となる．

確率分布の間の距離としてはいろいろな理由から Kullback-Leibler ダイバージェンス

$$D[p_{\boldsymbol{y}}(\boldsymbol{y}) \mid q_{\boldsymbol{y}}(\boldsymbol{y})] = E_{p_{\boldsymbol{y}}}[\log p_{\boldsymbol{y}}(\boldsymbol{y}) - \log q_{\boldsymbol{y}}(\boldsymbol{y})] \tag{15}$$

をとる[9]．

[8] これが主成分分析で，U, D を部分行列 $U = [U_1, U_2], D = diag[D_1, D_2]$ にわけてやると (サイズはうまくあわせる)，$U_1 D_1$ は Frobenius ノルムの意味で x の最良近似を与える部分空間への射影となる． また，データが多変量主成分分析に従うとき最大情報量抽出となっている．

[9] いろいろな距離がある中で KL を選ぶのは，一つは計算上の理由，もう一つは情報幾何的な意味をつけやすいという理由である．

$q_{\boldsymbol{y}}(\boldsymbol{y}) = \prod_{i=1}^{n} p_{y_i}(y_i)$ を取れば，独立性の規準として

$$L(W) = D[p_{\boldsymbol{y}}(\boldsymbol{y}) \mid \prod_{i=1}^{n} p_{y_i}(y_i)] \tag{16}$$

を最小化する W を見つける問題に帰着される．

この規準の難点は，確率密度関数に依存しているため，そのままだと密度関数の推定問題という厄介な問題を解かねばならない．実際には上記の問題を簡略化して解くことになる．

エントロピー最小化　後々のために上の式を少し変形しておく．上記の規準は，エントロピー関数 $H(\cdot) = -E[\log p_{\cdot}(\cdot)]$ を用いて書くと，

$$L(W) = \sum_{i=1}^{n} H(y_i) - H(\boldsymbol{y}) \tag{17}$$

となるが，$\boldsymbol{y} = W\boldsymbol{x}$ なので，確率変数の変換を考える．

最初に示したように線形変換した確率密度の間には

$$p_{\boldsymbol{y}}(\boldsymbol{y}) = \frac{p_{\boldsymbol{x}}(W^{-1}\boldsymbol{y})}{|\det W|} \tag{18}$$

の関係がある．

従って，

$$\begin{aligned} H(\boldsymbol{y}) &= -\int p_{\boldsymbol{y}}(\boldsymbol{y}) \log p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y} \\ &= -\int \frac{p_{\boldsymbol{x}}(W^{-1}\boldsymbol{y})}{|\det W|} \log \frac{p_{\boldsymbol{x}}(W^{-1}\boldsymbol{y})}{|\det W|} d\boldsymbol{y} \\ &= -\int p_{\boldsymbol{x}}(\boldsymbol{x})(\log p_{\boldsymbol{x}}(\boldsymbol{x}) - \log|\det W|) d\boldsymbol{x} \\ &= H(\boldsymbol{x}) + \log|\det W| \end{aligned} \tag{19}$$

つまり，

$$L(W) = \sum_{i=1}^{n} H(y_i) - \log|\det W| - H(\boldsymbol{x}) \tag{20}$$

であるが，最後の $H(\boldsymbol{x})$ は W によらない項なので，前２項だけが問題となる．仮に $|\det W| = 1$ の範囲内で探すとすると[10]，$L(W)$ の最小化は，個々の y_i 成分としてエントロピーの小さいものを選ぶようにするということになる．

[10] 自由度があるのでそれでも十分一般的である

エントロピーの近似　上の変形により，問題が各成分のエントロピーと $\det W$ の関数の和という関数の最適化問題に帰着された．

実際の信号が与えられたときには，そこからエントロピー関数を評価する必要があるが，その近似・推定手法によりいろいろなバリエーションができる．

データ点 $\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)$ だけが与えられたとき，統計的推定問題では，真の分布による平均をサンプル平均で近似することができる．つまり，エントロピーならば，

$$H(y_i) \simeq -E_T[\log p_{y_i}(y_i)] \tag{21}$$

となる． ただし，E_T は経験平均

$$E_T[f(y_i)] = \frac{1}{T}\sum_{t=1}^{T} f(y_i(t)) \tag{22}$$

ところがこの中の $\log p_{y_i}(y_i)$ も未知の関数だから，それをデータから計算可能な関数 $\log q_i(y_i)$ で近似してやる必要がある[11]．

素朴なのは分布を近似する方法であろう．主なものを以下に挙げる．

1. パラメトリックな推測
   パラメータ θ をもつ分布

   $$q_{y_i}(y_i) = f(y_i; \theta), \tag{23}$$

   をデータから推定して用いる．どういうクラスの分布を用いるかが問題となる．

2. ノンパラメトリックな推測
   いわゆるカーネル密度推定

   $$q_{y_i}(y_i) = E_T[f(y_i(t); \theta)] \tag{24}$$

   によって分布を推定する． カーネル関数 f として正規分布などを取るのが一般的だが，カーネル幅の決め方などの問題点がある．

[11] $q_i(y_i)$ はもはや分布である必要はない．この近似については推定関数とのからみでもう一度触れる．

3. 直交関数展開
   密度関数を正規分布のまわりで直交展開する．簡単のため y_i が平均 0, 分散 1 で正規化していると仮定すると，Gram-Charlier 展開は

$$q_i(y_i) = \phi(x)\left[1 + \sum_{k=3}^{\infty} \frac{\kappa_k}{k!} h_k(x)\right] \tag{25}$$

   となる． ただし，$h_k(x)$ は k 次エルミート多項式で，κ_k は k 次のキュムラント，つまり特性多項式[12]のテーラー展開の係数,

$$\log c_x(\omega) = \sum_{k=1}^{\infty} \frac{\kappa_k}{k!} (i\omega)^k. \tag{27}$$

   具体的には，平均 0 の確率変数 x に対して，$\kappa_1 = E[x]$, $\kappa_2 = E[x^2]$, $\kappa_3 = E[x^3]$, $\kappa_4 = E[x^4] - 3E[x^2]^2$ 等となる．

ここから派生したものとしてキュムラント κ_3, κ_4 の最大化 (最小化) が挙げられる．

- 平均と分散を固定したときにエントロピーが最も大きい確率分布は正規分布である．
- したがって，エントロピーの小さい分布は非正規性の強い分布ということになる．
- 正規分布の 3 次以上のキュムラントは 0 なので，これらの絶対値が大きくなるようにするのがよさそうである．
- また，中心極限定理から，独立な分布の重ねあわせが正規分布に近づくことからも正規分布から遠ざかるのが望ましいことが定性的にわかる．

4 その他の基準

高次相関 (含むカーネル正準相関)，定常性の仮定の元で相互相関など．

[12]特性多項式は密度関数のフーリエ変換.

$$c_x(\omega) = \int \exp(i\omega x) p_x(x) dx \tag{26}$$

5　最適化法

最適化問題はエントロピーの近似により,

$$L_q(W) = -\sum_{i=1}^{n} E_T[\log q_i(y_i)] - \log|\det W| \tag{28}$$

の最小化問題に帰着された．一般にこれを最適にする W は閉じた形で求められない.

最急降下法　そうした場合に用いられるのが勾配法 (最急降下法) である．(ユークリッド空間での) 勾配法とは，W の関数 $L_q(W)$ に対して，勾配 $\partial L_q(W)/\partial W$ を計算して，ある初期解からスタートし，勾配の反対方向に少しずつ解を改善させていく方法である.

$$W' = W - \epsilon\frac{\partial L_q(W)}{\partial W}. \tag{29}$$

ただし，これが最急降下方向になるのはユークリッド計量の場合だけである.

ここで，この微分についてもう少し詳しく見ておこう．$|\det W|$ とあるのは面倒なので，$\det W > 0$ と仮定する．スケールの自由度があるのでこう置いても一般性を失わない．さて，$\det W$ の余因子を D_{ij} とすると，任意の $j = 1, \ldots, n$ について

$$\det W = \sum_{i=1}^{n} w_{ij} D_{ij}, \tag{30}$$

となるので,

$$\nabla \log\det W = \frac{1}{\det W}\nabla \det W = \frac{1}{\det W}(D_{ij}) \tag{31}$$

である．ここで，$\nabla = (\partial/\partial w_{ij})$.

一方，$W^{-1} = (1/\det W)(D_{ji})$ だから,

$$\nabla \log\det W = W^{-\top} \tag{32}$$

ただし，$W^{-\top} = (W^{-1})^{\top}$ とおいた.

一方,

$$
\begin{aligned}
-\nabla \log q_i(y_i) &= -(\frac{\partial \log q_i(y_i)}{\partial w_{ij}}) = -(\frac{\partial \log q_i(\sum_{k=1}^n w_{ik}x_k)}{\partial w_{ij}}) \\
&= (-\frac{d \log q_i(y_i)}{dy_i} x_j) = \boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{x}^\top
\end{aligned} \tag{33}
$$

となる． ただし，$\phi_i(y) = -d \log q_i(y_i)/dy_i$. 結果的に,

$$
\nabla L_q(W) = E_T[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{x}^\top] - W^{-\top} = E_T[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^\top - I]W^{-\top} \tag{34}
$$

を得る.

自然勾配　以下簡単のため最適化する関数を f と書くことにする.

行列の空間を考えるときに，ユークリッド計量はあまり適切とはいえない．それは，異なる 2 点 W_1 と W_2 の間で計量が保存されないからである．適切なリーマン空間で計量を考慮した上で最も勾配の急な方向を,ユークリッド的な勾配

$$
\nabla f(W) = \left(\frac{\partial f(W)}{\partial w_{ij}} \right) \tag{35}
$$

に対比させて，自然勾配と呼び $\mathrm{grad}_W f(W)$ と書く.

Lie 群の不変計量に基づく最急降下法　ここでは，Lie 群の不変計量に基づく最急降下法を導く.

$W \in GL(n)$ は W^{-1} をかけることによって I に移される．点 W での接空間を $T_W GL(n)$ とする．$V_1, V_2 \in T_W GL(n)$ に対する計量は，V_1, V_2 を W^{-1} によって T_I に移した点での計量として定義すれば不変性が保たれる． T_I での計量は $\mathbb{R}^{n\times n}$ 上のユークリッド計量とすると,

$$
g(V_1, V_2) = \mathrm{tr}[W^{-\top} V_1^\top V_2 W^{-1}] = \mathrm{tr}[W^{-1} W^{-\top} V_1^\top V_2] \tag{36}
$$

となる.

最急降下方向は，計量行列 G としたときに,

$$
\mathrm{grad}_W f = G^{-1} \mathrm{vec}[\nabla f(W)] \tag{37}
$$

として得られる． が，これは vec とか入っていて少し厄介なので次のようにして求める．ここで，$V \in T_W GL(n)$ で，$\|V\|^2 = g(V, V) = c$ を満

たすものの中から最急降下方向を探す．最急降下法は $W' = W - \epsilon V$ とするので，

$$f(W') = f(W - \epsilon V) \tag{38}$$

を最小にする V を見つけるのだが，ϵ が小さいとして一次近似して考えればよく，

$$f(W') = -\epsilon \mathrm{tr}\left[\nabla f(W)^\top V\right] + o(\epsilon) \tag{39}$$

となるので，第一項だけを取り出すと，Lagrange の未定係数 λ を導入して，

$$L(V) = -\epsilon \mathrm{tr}\left[\nabla f(W)^\top V\right] - \lambda(c - \mathrm{tr}\left[W^{-1}W^{-\top}V^\top V\right]) \tag{40}$$

の停留点 V を求めればよい．

$$-\epsilon \nabla f(W)^\top + 2\lambda W^{-1}W^{-\top}V^\top = 0 \tag{41}$$

だから，

$$V = \frac{\epsilon}{2\lambda}\nabla f(W)W^\top W \tag{42}$$

となる． c を適当に取れば $\epsilon/(2\lambda) = 1$ としてよいので，

$$V = \nabla f(W)W^\top W \tag{43}$$

となる．$f(W) = L_q(W)$ として，式 (34) を代入すると，結局

$$\mathrm{grad}_W f(W) = E_T[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^\top - I]W \tag{44}$$

となる．不変な計量を導入することにより，面倒な逆行列の計算も不要となり一石二鳥となった．

直交群上の最適化　独立ならば無相関なので，前処理として主成分分析とスケーリングにより直交行列の自由度だけにして，一般の行列のかわりに，直交行列行列の空間だけで探すことを考えよう[13]．

[13]ただし，実際の信号は完全に独立にはならないので，その場合は最も独立な信号を探すことになる． その場合は損失関数を最小化するのが必ずしも無相関とは限らないので注意を要する．

ちなみにこのような無相関化の前処理のことを白色化 (whitening) とか sphering とかいう.

この場合も Lie 群の考え方に基づく最急降下法を考えることができる.

今 $W \in O(n)$ とすると，$T_W O(n)$ は，W を通るなめらかな曲線の速度ベクトルとして与えられるので,

$$W(t)^\top W(t) = I, \quad W(0) = W \tag{45}$$

を t で微分すると,

$$\dot{W}(0)^\top W + W^\top \dot{W}(0) = 0 \tag{46}$$

なので，$W(0) = I$ のときは,

$$\dot{W}(0)^\top + \dot{W}(0) = 0 \tag{47}$$

つまり反対称行列である． 従って，$V \in T_I O(n)$ なら V は反対称．逆に，任意の反対称行列 V について，$c(t) = (I + tV/2)(I - tV/2)^{-1}$ は $\dot{c}(0) = V$ を満たす $O(n)$ 上の曲線だから，$V \in T_I O(n)$. なお，$T_I O(n)$ は $SO(n)$ に付随する Lie 環 $\mathfrak{so}(n)$ である． W での接空間は $T_I O(n)$ の W による右移動または左移動で得られる.

さて，$T_W O(n)$ の計量はユークリッド空間から誘導される，$\mathrm{tr}[V_1^\top V_2]$ で定義する． この場合はこれで等長写像 (isometric) になっている．($GL(n)$ の場合の $W^{-1}W^{-\top}$ が単位行列になって消えるから)

また，I を $R^{n\times n}$ の元と見たとき，その接空間の元 $V \in T_I R^{n\times n}$ は $T_I O(n)$ と $T^\perp O(n)$ の直和で書けることに注意しておく.

$$V = \frac{V - V^\top}{2} + \frac{V + V^\top}{2}. \tag{48}$$

$O(n)$ の場合，自然勾配は次の定理に基づいて求める.

定理 3 多様体 (M, g) が $(\mathbb{R}^m, h)$ に埋め込まれているとし，g は h から誘導された計量とする．このとき $a \in M$ における f の自然勾配 $grad_a^M f$ は $grad_a^{\mathbb{R}^m} f$ を $T_a M$ に直交射影したものになる.

そこで，$O(n)$ は今 $\mathbb{R}^{n\times n}$ に埋め込まれており，その計量もユークリッド空間から誘導されたものであった.

次のような手続きで，ユークリッド勾配 ∇f を $T_W O(n)$ に直交射影する.

1. $W^\top$ をかけて $T_I\mathbb{R}^{n\times n}$ に左平行移動する.

$$\nabla f \mapsto W^\top \nabla f \tag{49}$$

2. $T_I\mathbb{R}^{n\times n}$ から式 (48) を用いて直交射影する.

$$W^\top \nabla f \mapsto \frac{1}{2}(W^\top \nabla f - \nabla f^\top W) \tag{50}$$

3. W をかけて $T_W O(n)$ に引き戻す.

$$\frac{1}{2}(W^\top \nabla f - \nabla f^\top W) \mapsto \frac{1}{2}(\nabla f - W\nabla f^\top W) \tag{51}$$

方向だけを問題にしているので簡単のため頭の $1/2$ を取ってやると,

$$\mathrm{grad}_W f = \nabla f - W\nabla f^\top W \tag{52}$$

となる.

測地線　さて，最急降下法では，$W' = W - \epsilon \mathrm{grad}_W f$ とするが，$O(n)$ の場合には一般に W' はもはや $O(n)$ の元ではない．そこで，その代わりに $\mathrm{grad}_W f$ を出発点とする測地線方向に動かすことを考える． $I \in O(n)$ から速度ベクトル X で出発する測地線は，単に行列指数関数

$$\psi(I, t, X) = \exp(tX) \tag{53}$$

で与えられるので，以下のようにして $W \in O(n)$ での測地線を得る.

1. まず $\mathrm{grad}_W f$ を $T_I O(n)$ に戻す.

$$\nabla f - W\nabla f^\top W \mapsto W^\top \nabla f - \nabla f^\top W \tag{54}$$

2. I での測地線を求める

$$W^\top \nabla f - \nabla f^\top W \mapsto \exp(t(W^\top \nabla f - \nabla f^\top W)) \tag{55}$$

3. W まで引き戻す.

$$\exp(t(W^\top \nabla f - \nabla f^\top W)) \mapsto W\exp(t(W^\top \nabla f - \nabla f^\top W)) \tag{56}$$

擬測地線　最近は技術も進歩したので，行列指数関数を計算することもたやすいが，それを近似する以下のパラメトリックな関数も興味深い．

$X \in \mathfrak{so}(n)$

$$\Theta_\alpha(X) = (I + \frac{X}{\alpha})^{\alpha/2}(I - \frac{X}{\alpha})^{-\alpha/2}, \quad \alpha \neq 0, \quad \alpha \in \mathbb{R} \tag{57}$$

- $\alpha = 1$ は対称直交化法 (つまり $I + X$ を固有展開して，固有値をすべて大きさ 1 に丸める方法)

$$\Theta_1(X) = \{(I + X)(I + X)^\top\}^{-1/2}(I + X) \tag{58}$$

- $\alpha = 2$ は Cayley 変換

$$\Theta_1(X) = (I + \frac{X}{2})(I - \frac{X}{2})^{-1} \tag{59}$$

- $\alpha \to \infty$ は指数関数

$$\Theta_\infty(X) = \exp(X) \tag{60}$$

- $\Theta_\alpha(tX)$ は t の 2 次のオーダーまで一致する．

$$\Theta_\alpha(tX) = I + tX + t^2\frac{X^2}{2} + O(t^3) \tag{61}$$

不動点法　FastICA

6　セミパラメトリック推定としての ICA

撹乱パラメータとセミパラメトリック推定　損失関数を見ると，本来知りたい A という行列のほかに，関数自由度の確率密度関数が入っている．推定する必要はないのだが，モデルに必然的に入ってしまうようなパラメータを撹乱パラメータという．また，このような関数自由度の撹乱パラメータが入っているような状況での統計的推定をセミパラメトリック推定という．

今，一般論を展開するために統計モデル $p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi})$ を考えよう ($\boldsymbol{\theta}$ が推定したいパラメータ，$\boldsymbol{\xi}$ が撹乱パラメータ)．

$$\boldsymbol{u}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi}) \tag{62}$$

$$\boldsymbol{v}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\xi}} p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\xi}) \tag{63}$$

を考え，このサンプル平均の停留点が最尤推定である．しかしながら，ナイーブにすべてのパラメータを最尤推定などで推定してしまうと，攪乱パラメータに余計な情報量を吸い取られてしまう[14]．攪乱パラメータが有限次元のときは，最尤推定に攪乱パラメータの影響はないが，攪乱パラメータが無限次元になるとそうはいかなくなる．

推定関数　このような状況下で，重要な役割を果たすのが推定関数というものの存在である．

推定関数 $\boldsymbol{z}(\boldsymbol{x},\boldsymbol{\theta})$ は $\boldsymbol{\xi}$ によらないベクトル値関数で，任意の $\boldsymbol{\theta},\boldsymbol{\xi}$ に対して，

$$E_{\boldsymbol{\theta},\boldsymbol{\xi}}[\boldsymbol{z}(\boldsymbol{x},\boldsymbol{\theta})] = \mathbf{0}, \tag{67}$$

$$\det K \neq 0, \qquad K = E_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{z}(\boldsymbol{x},\boldsymbol{\theta})\right] \tag{68}$$

$$E_{\boldsymbol{\theta},\boldsymbol{\xi}}[\boldsymbol{z}(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{z}(\boldsymbol{x},\boldsymbol{\theta})^\top] < \infty \tag{69}$$

を満たすものをいう[15]．このとき，

$$\sum_{t=1}^{T} \boldsymbol{z}(\boldsymbol{x}(t),\boldsymbol{\theta}) = 0 \tag{70}$$

の解 $\hat{\boldsymbol{\theta}}$ を M 推定量という．M 推定量は $T\to\infty$ で真の解に収束するという意味で一致性をもち，その推定値の分散は，漸近的に

$$\frac{1}{T}K^{-1}E_{\boldsymbol{\theta},\boldsymbol{\xi}}[\boldsymbol{z}\boldsymbol{z}^\top]K^{-\top} \tag{71}$$

[14]定量的には，Crameŕ-Rao の不等式

$$V \geq \frac{1}{n}\begin{bmatrix} G_u & G_{uv} \\ G_{vu} & G_v \end{bmatrix}^{-1} \tag{64}$$

を考えると，知りたいパラメータについては，攪乱パラメータを知らない場合は

$$V_\theta \geq \frac{1}{n}(G_u - G_{uv}G_v^{-1}G_{vu})^{-1} \tag{65}$$

で，知っている場合は

$$V_\theta^* \geq \frac{1}{n}G_u^{-1} \tag{66}$$

となるので，この差が攪乱パラメータによる損失である．

[15]$\boldsymbol{z}$ が推定関数ならそれに $\boldsymbol{\theta}$ だけに依存する任意の正則行列 $R(\boldsymbol{\theta})$ をかけたものはやはり推定関数である．

となる.

さて，ICA に戻って考えると，

$$F(\boldsymbol{y}, W) = I - \boldsymbol{\psi}(\boldsymbol{y}) \tag{72}$$

は適当に $\boldsymbol{y}$ の各成分をスケーリングすることにより推定関数となる[16]．F の ij 成分は $i \neq j$ のとき，独立なら真の分布がなんでも，正しい W をとれば

$$-E_{y_i,y_j}[\psi(y_i)y_j] = 0 \tag{73}$$

となるし，$i = j$ のときは

$$1 - E_{y_i}[\psi(y_i)y_i] \tag{74}$$

で，y_i のスケーリングは自由なので，$E_{y_i}[\psi(y_i)y_i]$ が 1 とすることは常にできるので，その場合 0 となる.

推定関数の中でどのようなものがいいかについては，情報幾何的な議論が必要.

漸近論

安定性

7　時間遅れのある場合

8　観測の次元が源信号の数と異なる場合

多い場合　Stiefel 多様体上の最適化

因子分析

少ない場合　スパースコーディング

謝辞

本稿は参考文献 [1, 2, 3] を全面的に参考にしておりますが，本稿の誤植・不明な点は赤穂の責任です.

[16] もちろん $\boldsymbol{\psi}$ は推定関数の条件を満たすものを取る必要がある.

参考文献

[1] 甘利俊一：独立成分分析とその周辺，In 甘利他 (編)：統計科学のフロンティア５：多変量解析の展開 隠れた構造と因果を推理する I, 岩波書店 (2002)

[2] A. Hyvärinen, J. Karhunen, E. Oja: Independent Component Analysis, John Wiley & Sons (2001)

[3] 村田昇：入門独立成分分析，東京電機大学出版局 (2004)